Recommender Systems—utilizing improved Collaborative Filtering Classifier as Machine Learning approach

A Thesis Submitted in Partial Fulfillment of the Requirement for The Degree of Bachelors of Science (Engineering) in Computer Science and Engineering

Submitted by:

Anisa Ibnat Sabrina

2018-1-17-014

Session: 2018-2022

Supervised to:

Mst. Shahnaj Parvin
Associate Professor and Chairperson

Department of Computer Science & Engineering

Central Women's University

APPROVAL

The thesis paper "Recommender Systems—utilizing improved Collaborative Filtering Classifier as Machine Learning approach" submitted by Anisa Ibnat Sabrina, 2018-1-17-014 to the Department of Computer Science and Engineering, Central Women's University, Dhaka Bangladesh, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Masters of Science in Computer Science and Engineering and approved as to its style and contents.

I found that they are very perseverant, skilled, and enthusiastic to accomplish this thesis within the expected time limit.

Supervisor:

……

Mst. Shahnaj Parvin

Associate Professor and Chairperson,

Department of Computer Science and Engineering

Central Women's University, Dhaka-1203.

DECLARATION

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by me under the supervision of Ayesha Aziz Prova, Senior Lecturer, Department of Computer Science and Engineering, Central Women's University, Dhaka, Bangladesh. We also declare that no part of this thesis and thereof has been or is being submitted elsewhere for the award of any degree or diploma.

Signature:

…….

Anisa Ibnat Sabrina

Candidate

Countersigned:

…….

(Mst. Shahnaj Parvin)

Supervisor

The research entitled “Recommender Systems—utilizing improved Collaborative Filtering Classifier as Machine Learning approach” submitted by Anisa Ibnat Sabrina, 2018-1-17-015, Session: Fall 2022 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering on 19 June 2022.

Board of Examiners

Mst. Shahnaj Parvin Chairperson

Associate Professor

Department of CSE, CWU

Ms. Sher Shermin Azmiri Khan (Internal Member)

Assistant Professor

Department of CSE, CWU

Ayesha Aziz Prova (Internal Member)

Senior Lecturer Supervisor

Department of CSE, CWU

Dr. Uzzal Kumar Acharjee (External Member)

Professor

Department of CSE

Jagannath University, Dhaka, Bangladesh

ACKNOWLEDGEMENT

In this very special moment, first and foremost We would like to express our heartiest gratitude to the almighty Allah for allowing us to accomplish this B.Sc. study successfully.

We are really thankful for the enormous blessing that the almighty has bestowed upon us not only during our study but also throughout our life.

In achieving the gigantic goal, we have gone through the interactions with and help from our supervisor and also other people. We would like to extend our deepest appreciation to those who have contributed to this dissertation itself in an essential way.

We would like to express our heartfelt thanks to all of us for being with us with immense support and care and to make this work success.

Anisa Ibnat Sabrina

2022

ABSTRACT

Recommender system has vital importance to users by providing more personalized information services. Its goal to forecast users interests by personalizing their likes or dislike. This is the platform where machine used to know anyone by personalizing him/her self or even neighbor, friends. Recommender system is one of the potential machine learning systems that could suggest anyone by their choices product or item. This system embeds information finding techniques to the problem of solving items recommendation during a need of users or clients. That is why the e-commers platform or any online suggestion to the customers or client is getting more easier than before. Recommender system is receiving worldwide success in online. There are three main classifiers for recommender system, they are content-based filtering, collaborative filtering and hybrid. In this paper, I work with collaborative filtering using a simple method that is Jaccard similarity and that has a great outcome. As I got 90% accuracy in the collaborative filtering using Jaccard similarity. In this proposed paper I will discuss about the other methods and distinguished with my approach.

Contents:

CHAPTER - 1

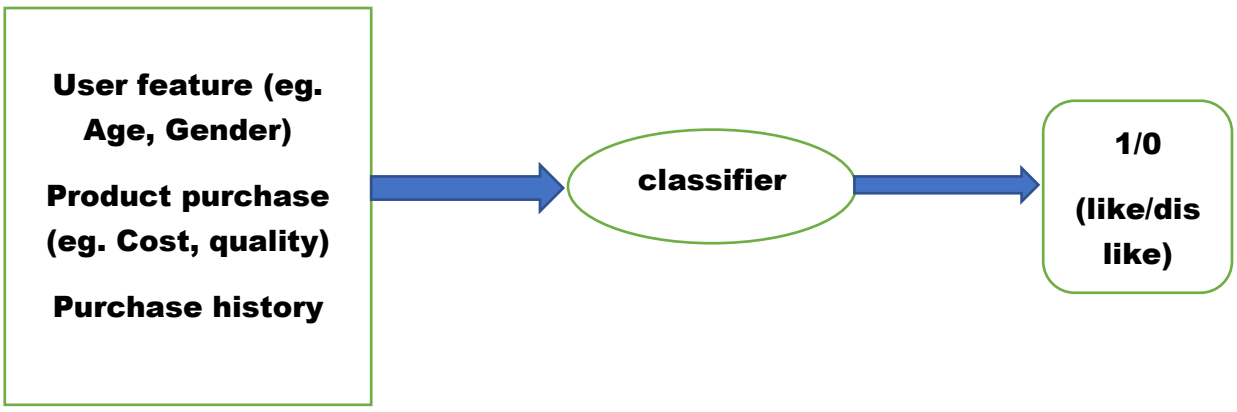
INTRODUCTION

1. INTRODUCTION

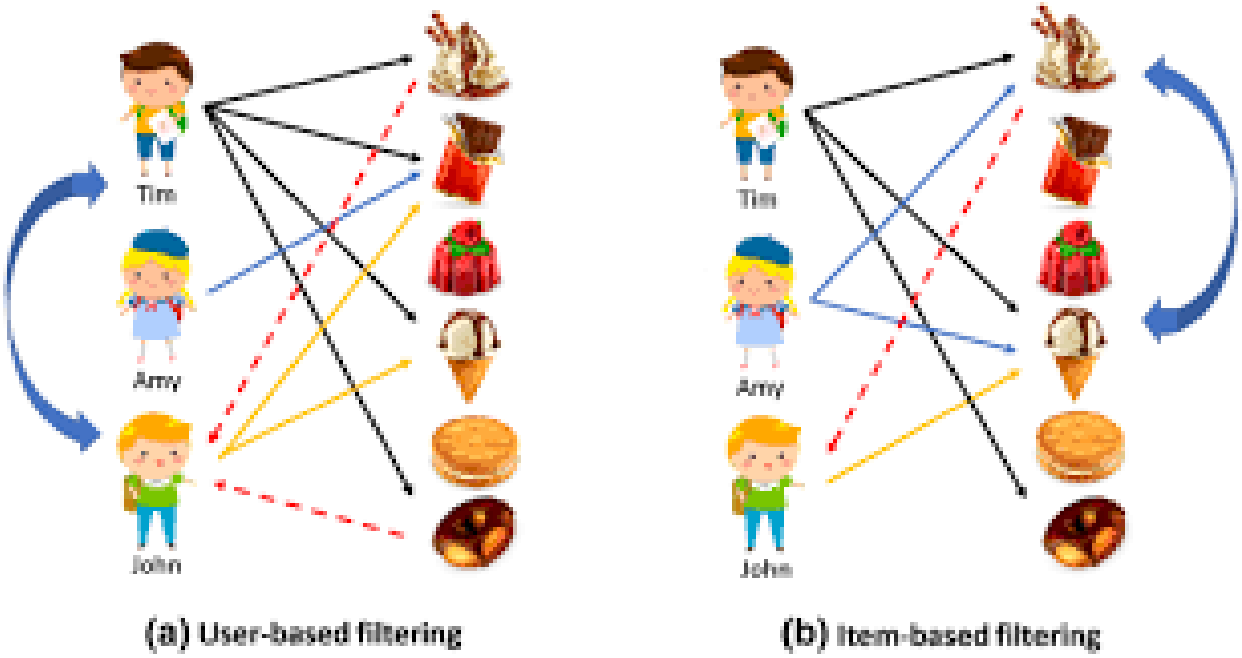
This chapter provide an overview of recommendation system using collaborative filtering. I discuss about several approaches of collaborating filtering and bring out the benefits and disadvantages of appointed methods. This paper clearly highlights the ideas behind the proposed method.

1.1 Overview

One of the highly uses artificial intelligence algorithm is recommender system which is usually combined with machine learning approach. Recommender system is immensely useful as they help users to discover products and services that users basically looking for. Basically, recommender systems are the systems that are aimed to suggest things to the user by personalizing content that based on many different factors. It predicts the most likely product that the users are most likely to collect and are of interest to. The recommender system deals with a huge volume of information by filtering the most essential information based on the data provided by a user and other thing is that it takes care of the user's attachment and interest. There are major three classifications for recommendation system, they are: 1. Content based filtering; 2. Collaborative filtering; 3. Hybrid recommendation approach. In a research paper [1], it shows there are eight categories in the recommendation field and data mining technique although the search was not enough, it serves as a comprehensive basis for understanding recommender systems research. Classification model that uses features of both products as well as users to predict whether a user will like a product or not.



A recommendation system is a subclass of Information filtering Systems that seeks to predict the rating or the preference a user might give to an item or product. Simply it is an algorithm that recommends which relevant with items to users. Like Netflix which movie to watch, in the case of e-commerce which product to buy, and another one is kindle which book to read, etc. The model that uses features of both products as well as users to predict whether a user will like a product or not. In this paper, I work with collaborative filtering. Collaborative filtering is making recommend according to combination of your experience and experiences of other people. There are two collaborative filtering methods: user-based CF and item-based CF.



(a) User-based filtering (b) Item-based filtering

The experimental analysis [2] explained collaborative filtering recommendation strategy is not established the user module, and directly the calculation is performed furthermore collaborative filtering algorithm based on the user model has the smallest error with the average number regardless of the number of user ratings. The collaborative filtering explains the recommender system very well and has a best way to personalized user by several users interests and preferences.

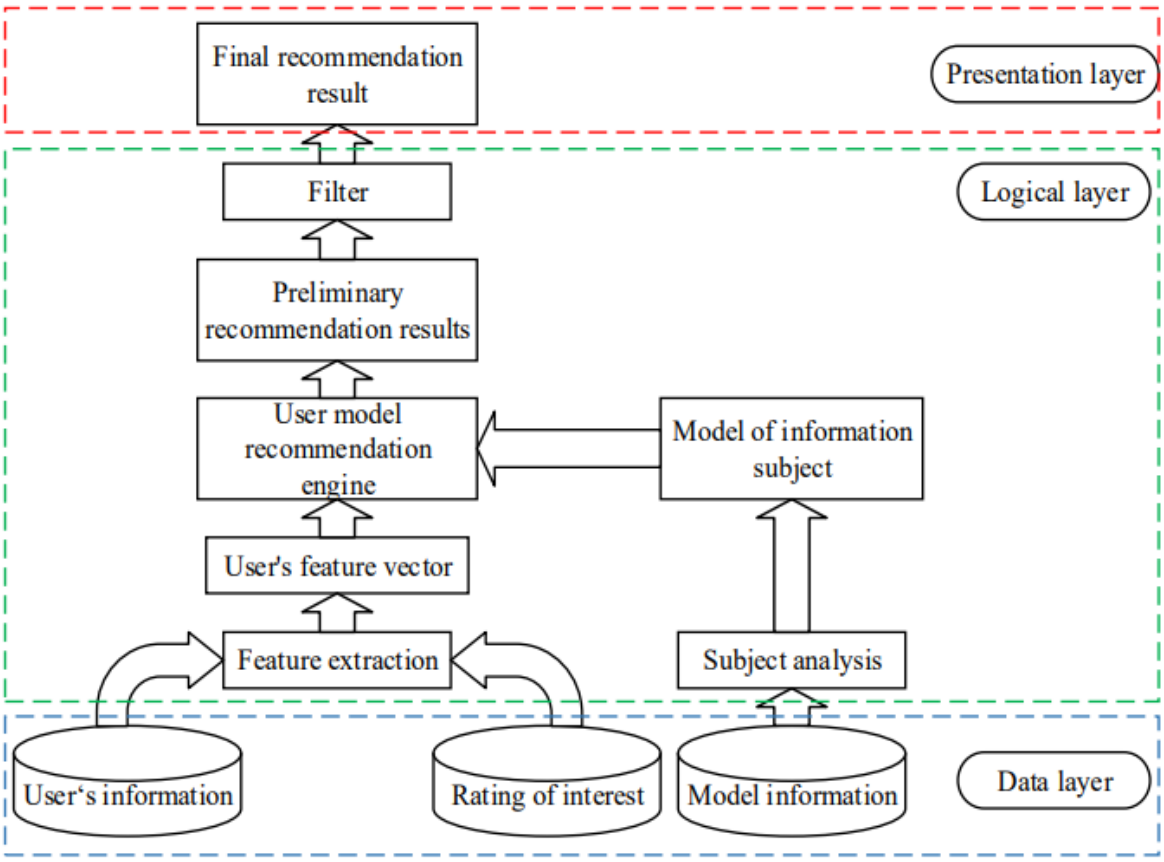


Figure 3. Recommended process analysis chart

There several methods to implement collaborative filtering. In my proposed paper I work with some method, I discuss and distinguished with their performances. In this paper, I used two types of datasets, one is movielens-100k and another is ACM recommender dataset.

1.2 Motivation

The recommendation system is established for 22 years which is an interdisciplinary discipline as it integrates information retrieval, prediction theory, function approximation theory and also nowadays, both the commercial field and the academical field are very concerned about the personalization of the recommendation system. It has a gigantic way to research and detect the limitation, implementation and improvement. Recommender system has many approaches but one has to find out the best possible features for particular activity. And that is why I choose to work with collaborative filtering and its methods to discuss and get the better accuracy of implementation.

1.3 Objectives

The objectives of my research is as follows:

- ✓ An implementation of recommender system.
- ✓ To get the better accuracy for the testing method.
- ✓ Find out the advantages and limitation of different methods.

1.4 Problem Description

Recommender systems aim to predict users' interests and recommend product items that quite likely are interesting for them. They are among the most powerful machine learning systems that online retailers implement in order to drive sales. Data required for recommender systems stems from explicit user ratings after watching a movie or listening to a song, from implicit search engine queries and purchase histories, or from other knowledge about the users/items themselves. An increasing number of online companies are utilizing recommendation systems to increase user interaction enrich shopping potential. Use cases of recommendation systems have been expanding rapidly across many aspects of eCommerce and online. E-commerce and retail companies are utilizing the power of data to boost sales with the help of recommender systems (RS) implemented on their websites. The use cases of these systems have been incrementing consistently.

There could be no better time than now to dive deeper into this excellent machine learning technique. Then thus I decide to work on it. Want to develop a method to implement recommender system. Jaccard coefficient has a limitation that it neglects the value of ratings and only considers the ratio of co-rated items. Hence I decide to work with the absolute value and detect the similarity between users who rated the same product of the dataset.

1.5 Organization of Thesis

The first chapter of my thesis illuminates the context, motivation, objectives, and problem description. Chapter (2) covers the Literature Review of collaborative filtering and their methods including. Different methods are briefly described in Chapter (3). Chapter (4) illustrates the recommended method. Chapter (5) illustrates the Experiment and result discussion. Chapter (6) gives the conclusion and future work. Chapter 7 is a list of the reference work.

CHAPTER - 2

LITERATURE REVIEW

2. LITERATURE REVIEW

Mubbashir Ayub, Mustansar Ali Ghazanfar, Tasawer Khan, Asjad Saleem [3] they have discussed several similarity measures for collaborative filtering. Jaccard similarity is one of the methods to measure the similarity item-to-item or user-to-user in collaborative filtering (CF). Jaccard similarity basically neglects absolute values of ratings and the average rating value of a user so they proposed an advanced performance which considers the ratio between absolute rating values

and number of commonly rated items. So they have categorized such users by rating performance. Their standered deviation calculation is:$RPB_{(a,b)} = cos(|\overline{R_a} - \overline{R_b}| \cdot |\sigma_a - \sigma_b|)$

And after calculated they got the final improved jaccard measure:

$$rating\_jaccard\_RPB = \left(\frac{|N_T(a,b)|}{|I_a \cap I_b|}\right) \cdot cos(|\overline{R_a} - \overline{R_b}| \cdot |\sigma_a - \sigma_b|)$$

Lamis Al Hassanieh, Chadi Abou Jaoudeh, Jacques Bou Abdo, Jacques Demerjian [4]worked with seven similarities for implementing collaborative filtering and made the comparison between them in same dataset. They showed the performance of the seven similarities using the prediction accuracy metrics (MAE, RMSE). They observed that pearson comparison similarity (PCS) has least accuracy then mean square difference (MSD). Frequency-weighted Pearson Correlation (FPC) and Weighted-Pearson Correlation (WPC) had higher accuracy in the 5% dataset and Cosine Vector Similarity (CVS) and Spearman Rank Correlation (SRC) showed better performances in the final 10% dataset.

Manolis Vozalis, Angelos Markos, Konstantinos Margaritis used three collaborative filtering (CF) method which rely on the singular value decomposition (SVD) of a perfectly transformed user-item ratings matrix. Mean Absolute Error (MAE) was the metric they [5] calculated to evaluate the accuracy of the methods and MAE

measures the predictions created by the recommender systems from the true rating values as specified by the user.

Bing Tang , Linyao Kang, Li Zhang , Feiyan Guo , and Haiwu [6] introduced NMF-based collaborative filtering on the platform effectively outperforms to consecutive user-based and item-based collaborative filtering with a higher processing speed and higher recommendation accuracy. The main problem of NMF is that the original matrix is makes the computational complexity very high. They explained that the heterogeneous CPU/GPU cluster, nodes have large memory resources and GPU multicore resources, and the advantages of distributed storage so a GPU-accelerated NMF algorithm on Spark platform has been designed to resolve the matter of low processing speed of NMF as the size of the matrix enhancement.

CHAPTER – 3

BASIC CONCEPT of RECOMMENDER SYSTEM

3.BASIC CONCEPT OF RECOMMENDER SYSTEM

Recommender systems (RS) are the systems that are intended to suggest things to the user based on many various factors. RS predict the most likely product that the users are most likely interested or purchased. RS deals with a large volume of information present by filtering the most important information based on the data provided by a user and other factors that take care of the user's interest. It finds out the match between user and item and imputes the similarities between users and items for recommendation [7]. Apparently, RS are defined as the supporting systems which help users to find information, products, or services by aggregating and analyzing suggestions from other users that is reviews from various authorities or other users [1] [2] [6]. RS have spread interpretatively over the world widely and its applications have growing over several territory of life [8].

The recommender systems can be categorized on several bases. In the literature, the categorization of the recommender systems is usually found [9] on the following bases;

♦ Approaches used

♦ Area of application for which recommendation is made

♦ Data mining techniques applied, etc.

3.1. Why the Recommendation System?

- Benefits users in finding items of their interest [1] [4] [6].
- Help item providers in delivering their items to the right user.
- Identity products that are most incidental to users [10].
- Personalized content with their friends and neighbor.
- Assistance of websites to enhance user engagement.

3.1.1 Classification of RS

Apparently, there are six types of recommender systems which work primarily in the Media and Entertainment industry:

➢ Collaborative Recommender system (CFS),

- Content-based recommender system,
- Demographic based recommender system,
- Utility based recommender system,
- Knowledge based recommender system,
- Hybrid recommender system.

In my proposed paper I work with collaborative filtering (CF).

3.1.2 Collaborative Filtering (CF)

Collaborative Filtering [11] is the process of filtering or evaluating items using the opinions of other users. CF filters information by using the interactions and data collected by the system from other users. It's based on the idea that user who agreed in their evaluation of items are likely to agree again in the future. CF systems conduct a database for a user preference to predict additional items or products a new user could choice [12].

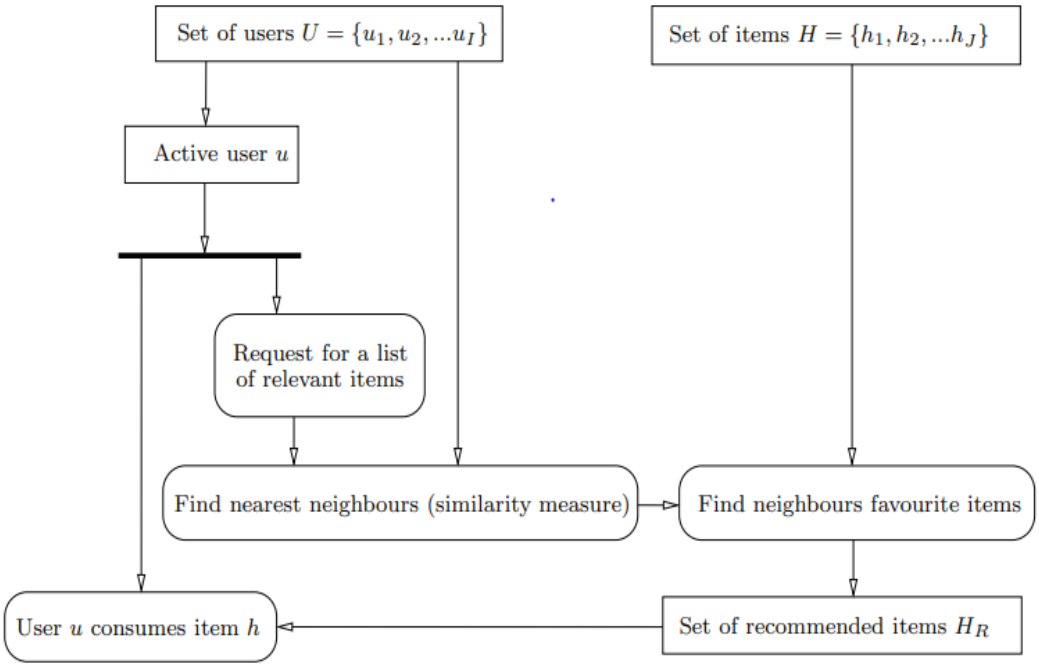


Figure: Collaborative Filtering recommendation system

3.1.3 Types of Collaborative Filtering

There are 2 types of basic Collaborative Filtering techniques are:

- **User-based Collaborative Filtering:** It proceeds results from implicit data; it is able to retrieve information that users otherwise might not provide. The first class of CF is the user-based approach.
- **Item-based Collaborative Filtering:** Item-based CF refers to the recommendation of items or products using collaborative filtering.

3.1.4 Types of Approaches for Collaborative Filtering

There are five types of approaches can use in collaborative filtering techniques:

- **Memory based**: Memory-Based CF approaches can be separated into two main parts: user-item filtering and item-item filtering. A user-item filtering takes a particular people, find users that are similar to that user based on similarity of ratings, and recommend items that those similar peoples liked.
- **Model based:** CF models are developed using machine learning algorithms to predict people's rating of unrated items.
- **Matrix Factorization:** The motive behind such models is that preferences of a user can be determined by a small number of hidden factors. Matrix factorization can be done by different methods and there are several research papers.
- **Clustering:** Clustering techniques [13] are commonly used for collaborative filtering recommendation.

- **Deep Learning:** There is a ton of research material on CF using matrix factorization or similarity matrix. But there is lack of components to learn how to use deep learning models for collaborative filtering.

3.2 Working Procedure

The methodology of this work includes following steps –

3.2.1 Data collection

I use MovieLens 100k dataset from Kaggle.com.

3.2.2 Data Preprocessing and Cleaning

In the dataset there have a huge amount of data around. And raw data is harsh. So it is necessary to do preprocessing and cleaning data. To process the data some steps are follows: The null values are drop. By dropping irrelevant and categorical attributes are converted into numerical values. The date attribute is splitting for the model.

3.2.3 Extraction of Features

Building the model, features selection plays vital role. I ignore timestamp feature.

3.2.4 Training Model and Fit Data

After Extraction of Features, the dataset divides dataset into training and test data randomly in ratio of 75:25. Then fit training data into proposed algorithm so that computer can get trained using this data. Now the training part is complete.

3.3 Data Mining Concepts

Data mining is the technique that helps to extract information from large sets of data and to identify trends, patterns, and important data. It is the process of examination and analysis of huge quantities of data in order to search significant rules and patterns. It is a subfield of computer science that is the technique of computational practice of huge data [14] [15]. Data collect and storing helps to accumulate large amounts of data at lower cost. Exploiting data storage, simultaneously extract useful information is the entire aim of the generic activity of data mining. There are several papers discusses the data mining techniques, algorithms that shows data mining technology to improved [16].

3.3.1 Fundamental steps in data mining

There are two types of data mining: descriptive mining and predictive mining with different functions and technologies [17]. The process of data mining is divided into two parts. Data Preprocessing and Data Mining. Data Preprocessing involves with the data cleaning, data integration, data reduction, and data transformation. The data mining part performs data mining, pattern evaluation and knowledge representation. The procedure of Data mining steps are in below :

- Data Cleaning: Remove all unclear, irrelevant, noisy, incomplete, null data from the collection.
- Data Integration: Multiple heterogeneous data sources are combined for analysis. It helps to improve the accuracy and speed of the model.
- Data Reduction: By reducing some data one may work with their data with pleasure.
- Data Transformation: The transformation of data is the process of transferring data from one format to another. Data integration and data management and data transformation is important.

- ❖ Data Mining: Data Mining is a process to place interesting patterns and knowledge from a huge amount of data.
- ❖ Pattern Evaluation: It describes patterns of information based on different types of steps. Pattern is potentially useful, understandable by humans and viewing to make data understandable.
- ❖ Knowledge Representation: The representation of knowledge is a indicate of sense to the user in terms of trees, tables, rules, graphs etc.

3.4 Data Types That Can Be Minded

The data mining method helps to analyze huge amounts of data quickly. To get desire purpose, data mining method can be applied for any type of appropriate data .

- ➢ The data from data warehouse in which data store is collected and integrated from one or more sources.
- ➢ The Data from external sources and databases.
- ➢ The data from the Spatial Database where data stores in the form of coordinates, lines, and different shapes, etc.
- ➢ The data from transactional database where data stores record that are captured as transactions.
- ➢ The data from the Flat files are in the form of binary.

3.5 Technique for Data Mining

Data mining techniques have been developed many technologies, including association, classification, clustering, prediction, database and data warehousing systems, Machine Learning, Neural Networks so on.

3.5.1. Classification:

It is the process of finding a model that describes and distinguishes data classes and concepts. In classification I build a model where I use machine learning approach as collaborative filtering.

3.5.2 Regression:

Regression technique predict the value of continuous valued variable which is basically predictive variable. Regression analysis is another data mining method of identifying and analyzing the relationship between variables by calculating predicted data values based on variables of the dataset [18].

3.5.3. Prediction:

This technique predicts the relationship that exists between independent and dependent variables. It analyzes past events or instances in a right sequence for predicting a future.

3.5.4. Machine learning:

Machine learning is the process by which computers use algorithms to learn and develop their performance depending on data. In data mining, machine learning's applications are vast. There are 4 methods for machine learning are as follows:

- Supervised learning: Supervised machine learning creates a model that makes predictions based on in a particular class of dataset.[4]
- Unsupervised learning: Unsupervised machine learning creates a model that makes predictions based on unknown class of dataset. Unsupervised models are used to perform clustering and association.[4]
- Semi-supervised learning: Semi-supervised learning uses a combination of labeled and unlabeled data, making it a hybrid of the above models. [4]

- Reinforcement learning: This is a more layered process in which computers learn to make decisions based on examining data in a specific environment.

3.6 Applications of Data Mining

The applications of data mining are given below:

- ❖ Detect the feature of the model.
- ❖ Use of visualization tools of recommender system for data analysis.
- ❖ Identify successful recommendation for different users by predicting ratings.
- ❖ Manufacturing Engineering.
- ❖ The application of user correlation in collaborative filtering [19].
- ❖ Detect security violations.

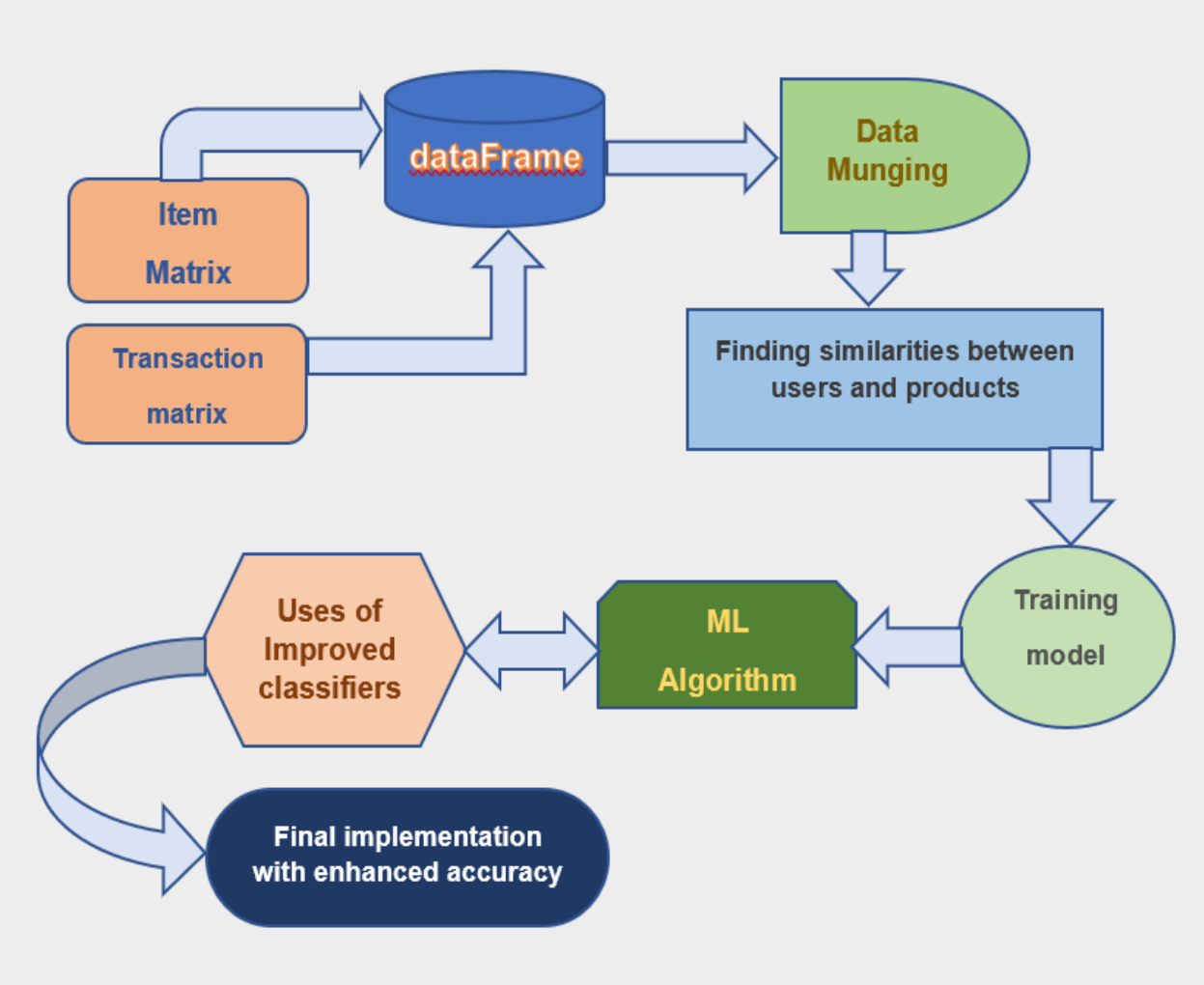
CHAPTER – 4

PROPOSED METHOD

4. PROPOSED METHOD

In this section, I describe methodology, dataset, data processing, experiment environments, and classification algorithm.

4.1 Methodology



Here is the methodology of my proposed experimental work. This experience is based on machine learning and data mining where I used some methods to improve the machine learning approach as collaborative filtering. The purpose of this paper, is to analyze collaborative filtering using jaccard similarity and improve the accuracy.

4.1.1. Experiments dataset

The dataset has used in this experiment is MovieLens 100k dataset where 100k ratings of approximately 9000 movies by 700 users.

	userId	movieId	rating	title	genres
0	1	1	4.0	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
215	1	3	4.0	Grumpier Old Men (1995)	Comedy|Romance
267	1	6	4.0	Heat (1995)	Action|Crime|Thriller
369	1	47	5.0	Seven (a.k.a. Se7en) (1995)	Mystery|Thriller
572	1	50	5.0	Usual Suspects, The (1995)	Crime|Mystery|Thriller
...	...	...	...	...	...
16056	1	3744	4.0	Shaft (2000)	Action|Crime|Thriller
16075	1	3793	5.0	X-Men (2000)	Action|Adventure|Sci-Fi
16208	1	3809	4.0	What About Bob? (1991)	Comedy
16243	1	4006	4.0	Transformers: The Movie (1986)	Adventure|Animation|Children|Sci-Fi
16250	1	5060	5.0	M*A*S*H (a.k.a. MASH) (1970)	Comedy|Drama|War

232 rows × 5 columns

Figure 1: Dataset of the experience

In the dataset there was two different csv file as ratings and movies. For convenience, I merged these to dataset into one dataset. The features of movies dataset are userId, movieId, rating, timestamp. And movies dataset has movieId, title, genres.

Table 1: Explanations of features

Attributes	Description
userId	Every user is represented by a unique Id.
movieId	Every movie is represented by an uniue Id.
rating	The rating given by the user to the corresponding movie.

timestamp	The time at which the rating was recorded.
genres	Represents category of the movie.
title	Movie name which is represented by the corresponding movieId.

Here, rating is count from 0 to 5 and average rating is 3.5 that is the mean ratings given by the users to all the movies. In this dataset I worked with 232 rows and 5 columns. In this work, 75% of the dataset was used for training and the remaining 25% was used for testing to improved proposed model classification's performance.

4.1.2. Data Preprocessing

In the dataset raw data are harsh so it is necessary to do preprocessing and cleaning data or data munging. Steps of data processing are follows:

i. For convenience, I ignored timestamp column of the dataset.
ii. Combined ratings and movies CSV file.
iii. The null values are dropped.
iv. Some impertinent attributes are also drop.

4.1.3. Experiment Environment

- Processor: 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz.
- Memory: 8.00 GB (7.73 GB usable).
- System type: 64-bit operating system, x64-based processor.
- Operating system: Windows 10
- Languages: Python
- Platform: Anaconda 3 (64 bit)

4.1.4. Feature Selection and Training Model

The Features selection is that can be used to build the model of the proposed. Enhanced public availability of ratings datasets will enable more effective research into CF system [11]. I used the columns of dataset for feature selection are userId, movieId, rating, timestamp, title, genres. To train the model, I split the dataset into training and testing dataset where the training dataset is used to fit the model, and the test dataset is used to evaluate the model. The data split in the ratio of 75% for training and 25% for testing.

4.1.5. Classification Algorithms

In this proposed paper I have work with collaborative filtering using jaccard similarity. Measuring the Jaccard similarity coefficient between two data sets is the result of division between the number of features [20].

$$Jaccard\_similarity(a,b) = (p(a \cap b) | p(a \cup b)) \quad \ldots\ldots\ldots\ldots (2)$$

In a research paper [21] they showed and calculated the similarity of 10878 pairs of the research fields for 2010-2019 and for 2000-2009 on base of dimensions data and the similarity of the research fields assessing with Jaccard index similarity. In another research they detect and analyze abusive YouTube comment by using jaccard similarity. Another thing is quite important to analyze my experience is k_fold cross validation is a general training/testing scenario for measuring the performance of a RS as well as a classifier [22]. There is a limitation in jaccard similarity that is it ignores both absolute values of ratings and the average rating value of a user but some recent works [3] [23] find out the solution and showed their experiences that considers the ratio between absolute rating values and number of commonly rated items or product.

4.1.6. Proposed algorithm of CF using jaccard similarity

1) Collect data.
2) Data Pre-processing
3) Assigning feature variables
4) Trained the model. (Improved jaccard similarity)

   Steps:

   - Select number of item, n and users similar users, k as the threshold value.
   - Find the number of users from train dataset.
   - Find number of similar users from train data using eq (2).
   - Then take k_users of number of minimum threshold users from similar users.
   - Then find top users from k_users.
   - Then find the top item among top users.
   - Then predict the list of items.

CHAPTER – 5

EXPERIMENT AND RESULTS

DISCUSSION

5. Experiments and Results Discussion

The performance measure is most vital part show the classifiers' activity. In a research paper [24] says the performance evaluation metrics used for the analysis of different RS that the set of research works, 35% of the works use recall measure, 16% of the works apply mean absolute error, 11% of the works take root mean square error, 41% of the papers consider precision, 30% of the contributions analyses F1-measure, 31% of the works apply accuracy and 6% of the works employ coverage measure to validate the performance of the RS. Besides, some additional dimensions are also used for validating the performance in a few applications. In this paper, I measure my model with accuracy, RMSE, and precision.

Accuracy

It is defined as the percentage of correct predictions out of all the observations

Accuracy = Correct Predictions / Total Cases * 100%

Accuracy = (TP + TN) / (TP + TN + FP + FN +) * 100%

Precision

It is defined as the percentage of true positive cases versus all the cases where the prediction is true.

Precision = True Positive / All Predicted Positives * 100%

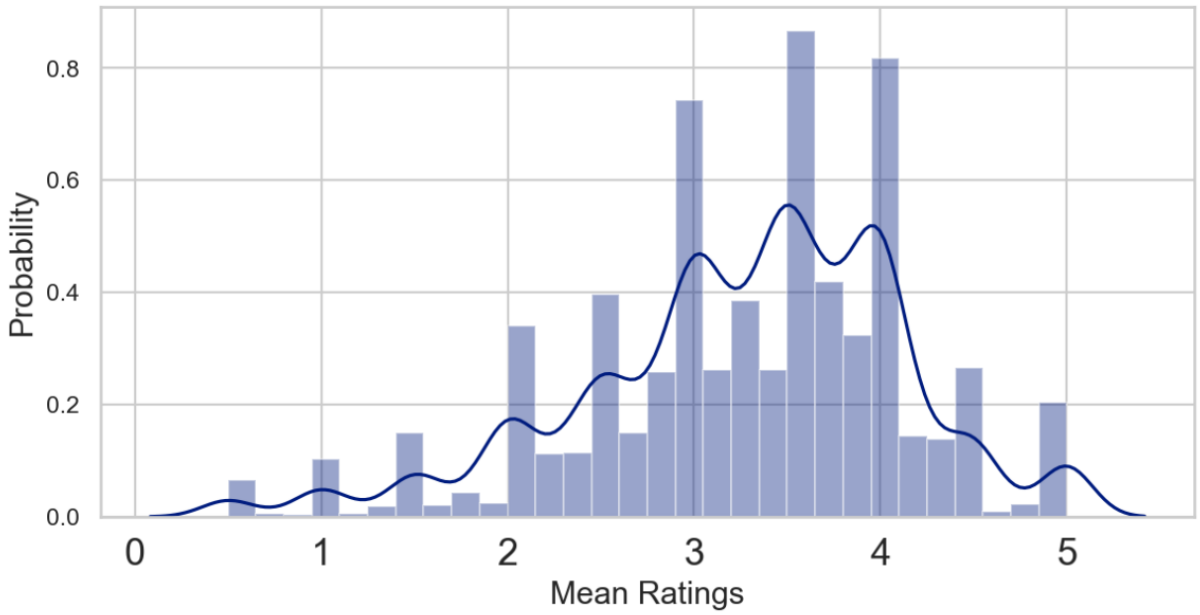
Precision = TP / (TP + FP) * 100%

RMSE

Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general purpose error metric for numerical predictions.

5.1 Optimum Similarity Measurement

5.1.1. Probability of rating



Here the graph shows that the number rating 3.5 is most probably get ratings from u sers. Total no of users that gave rating of 5.0 is 296 and total no of individual users that gave rating of 5.0 is 289.

5.1.2. RMSE comparing of methods

Mean square distance similarity

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9467	0.9461	0.9441	0.9442	0.9317	0.9426	0.0055
Fit time	0.40	0.40	0.39	0.38	0.33	0.38	0.03
Test time	1.81	1.93	2.14	1.73	1.95	1.91	0.14

Singular value decomposition

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.8740	0.8713	0.8642	0.8853	0.8725	0.8735	0.0068
Fit time	9.37	9.03	8.97	9.19	9.03	9.12	0.15
Test time	0.27	0.20	0.23	0.39	0.24	0.27	0.06

Non-negative matrix factorization

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9192	0.9295	0.9236	0.9239	0.9210	0.9235	0.0035
Fit time	10.23	10.25	10.02	9.63	10.35	10.09	0.26
Test time	0.30	0.31	0.28	0.17	0.33	0.28	0.06

Jaccard similarity

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.8731	0.8698	0.8687	0.8761	0.8755	0.8726	0.0030
Fit time	5.98	5.82	5.64	5.69	5.59	5.74	0.14
Test time	0.17	0.16	0.16	0.23	0.14	0.17	0.03

So as the lowest RMSE shows the result to measure the performance. In jaccard similarity there is 0.86 RMSE score which is the lowest then other methods.

5.1.3. Comparing accuracy from previous method

```
UHR   = 0.820
MAP   = 0.171
```

```
UHR   = 0.907
MAP   = 0.181
```

And in measuring accuracy I got 90% accuracy of my improved proposed method which was previously 82%. And the precision rate has also increased, so the mean average precision is 18%.

CHAPTER – 6

CONCLUSION AND FUTURE WORK

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion and limitation

RS have been a vital in platform on the web for the consumer to suggest items what they would be prefer. Number of users and items are incrementing, RS appoint the main shortcoming: data sparsity and data scalability issues, which bring out the poor quality of prediction and the inefficient time consuming. In practical, RS shorten transaction charge of searching and selecting items in an online shopping environment. Developing a recommendation system takes a significant understanding of data.

There is various collaborative based recommender system are proposed and explained in different method and approach. There are some papers are measured collaborative filtering using jaccard similarity but that is quite complex because they used various way of jaccrad index and that's why there RMSE score is not good. So, I prefer jaccard similarity in a simple technique and getting better accuracy. In my work I only implement the classifier in one dataset.

6.2. Future work

In future I would perform in several dataset to implement my improved algorithm of the classifier. And would like to work with naïve bayes to measure the trustworthiness of the recommendation.

CHAPTER – 7

REFERENCES

References

[1] H. K. K. I. Y. C. J. K. K. Deuk Hee Park, "A Review and Classification of Recommender Systems Research," *International Conference on Social Science and Humanity,* pp. 290-294, 2011.

[2] Y. G. a. X.-M. L. Bo Song, "Research on Collaborative Filtering Recommendation Algorithm Based on Mahout and User Model," *International Symposium on Big Data and Applied Statistics,* 2019.

[3] M. A. G. T. K. A. S. Mubbashir Ayub, "An Efective Model for Jaccard Coefcient to Increase the Performance of Collaborative Filtering," *Arabian Journal for Science and Engineering,* 2020.

[4] C. A. J. J. B. A. J. D. Lamis Al Hassanieh, "Similarity measures for collaborative filtering recommender systems," *IEEE Middle East and North Africa Communications,* 2018.

[5] A. M. K. M. Manolis Vozalis, "Evaluation of standard SVD-based techniques for Collaborative Filtering," *9th Hellenic European Research on Computer Mathematics and its Applications Conference,* 2009.

[6] L. K. L. Z. ,. F. G. ,. a. H. Bing Tang, "Collaborative Filtering Recommendation Using Nonnegative Matrix Factorization in GPU-Accelerated Spark Platform," *Hindawi,* vol. 2021, p. 15, 2020.

[7] D. W. M. M. W. W. G. Z. Jie Lu, "Recommender system application developments: A survey," *Decision Support Systems 74 (2015) 12–32,* p. 12–32, 2015.

[8] D. C. J. A. K. a. J. R. Mark O'Connor, "PolyLens: A Recommender System for," *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work, 16-20 September 2001, Bonn, Germany, Kluwer Academic Publishers. Printed in the Netherlands.,* pp. 199-218, 2001.

[9] J. S. a. R. A. Shahab Saquib Sohail, "Classifications of Recommender Systems: A review," *Journal of Engineering Science and Technology Review 10 (4) (2017) ,* pp. 132-153, 2017.

[10] M. D. G. A. F. P. L. G. S. F. R. LI CHEN, "Human Decision Making and Recommender System," *ACM Transactions on Interactive Intelligent Systems,* vol. 3, no. 3, p. 7, 2013.

[11] D. F. J. H. S. S. J. Ben Schafer, "Collaborative Filtering Recommender Systems," *P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): The Adaptive Web, Springer-Verlag Berlin Heidelberg,* p. 291 – 324, 2007.

[12] D. H. C. K. John S. Breese, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Appears in Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence,* 2013.

[13] C. H. Chih-Fong Tsai a, "Cluster ensembles in collaborative filtering recommendation," *Elsevier B.V. All rights reserved, Applied Soft Computing,* pp. 1417-1425, 2011.

[14] P. G. Sony V Hovale, "Survey Paper on Recommendation System using Data Mining technique," *International Journal Of Engineering And Computer Science ISSN: 2319-7242 ,* vol. 5, no. 5, pp. 16697-16699, 2016.

[15] A. M. P. S. S. Srinivasa G, "Survey Paper on Recommendation System using Data Mining Techniques," *International Journal of Engineering and Technical Research (IJETR),* vol. 6, pp. 2454-4698, 2016.

[16] M. B. M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," *Bharati M. Ramageri / Indian Journal of Computer Science and Engineering ,* vol. 1, no. 4, pp. 301-305 .

[17] S. N. A. Mustafa Abdalrassual Jassim, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," *IOP Conf. Series: Materials Science and Engineering ,* 2020.

[18] M. Y. R. Koti Neha, "A Study On Applications Of Data Mining," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH ,* vol. 9, no. 2, pp. 2277-8616, 2020.

[19] J. B. Schafer, "The Application of Data-Mining to Recommender Systems," *Encyclopedia of Data Warehousing and Mining, Second Edition,* p. 6, 2009.

[20] J. S. E. N. a. S. W. Suphakit Niwattanakul, "Using of Jaccard Coefficient for Keywords Similarity," *Proceedings of the International MultiConference of Engineers and Computer Scientists ,* vol. Vol I, 2013.

[21] S. S. a. M. Petrychko, "Jaccard index-Based Assessing the Similarity of Research Fields in Dimensions," *Creative Commons License Attribution 4.0 Internationa,* pp. Vol-2533, 2019.

[22] Z.-H. Z. X.-L. D. H.-R. Z. T.-J. L. L. Z. a. F. M. Shuang-Bo Sun, "Integrating Triangle and Jaccard similarities for recommendation," *PLoS ONE 12(8): e0183570,* 2017.

[23] Y. Lee, "RECOMMEND RECOMMENDATION SYSTEM USING COLLABORATIVE FILTERING," *Master's Theses and Graduate Research at SJSU ScholarWorks,* 2015.

[24] D. R. a. M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data, Article number: 59,* 2022.